Global optimisation for improved volume tracking of time-varying meshes

Jan Dvořák[0000−0003−4569−1151], Filip Hácha[0000−0001−8956−6411], and Libor Váša[0000−0002−0213−3769]

Department of Computer Science and Engineering, University of West Bohemia, Pilsen, Czech Republic {jdvorak,hachaf,lvasa}@kiv.zcu.cz

Abstract. Processing of deforming shapes represented by sequences of triangle meshes with connectivity varying in time is difficult, because of the lack of temporal correspondence information, which makes it hard to exploit the temporal coherence. Establishing surface correspondence is not an easy task either, especially since some surface patches may have no corresponding counterpart in some frames, due to self-contact. Previously, it has been shown that establishing sparse correspondence via tracking volume elements might be feasible, however, previous methods suffer from severe drawbacks, which lead to tracking artifacts that compromise the applicability of the results. In this paper, we propose a new, temporally global optimisation step, which allows to improve the intermediate results obtained via forward tracking. Together with an improved formulation of volume element affinity and a robust means of identifying and removing tracking irregularities, the procedure yields a substantially better model of temporal volume correspondence.

Keywords: Time-varying mesh · model · animation · tracking · analysis · surface.

1 Introduction

Sequences of triangle meshes are becoming more common in computer graphics due to a recent boom in both image acquisition hardware and reconstruction techniques aimed at estimating a static shape from a set of views. Since the most common data source is a reconstruction from video sequences, which treats each frame as an independent reconstruction problem, the most common type of resulting mesh sequences is the *Time-Varying Mesh (TVM)*, i.e. a sequence, where both the geometry (vertex coordinates) and the connectivity (triangles/polygons) are different in each frame. Effort has been put previously into converting this data into a more convenient form, e.g. a *dynamic mesh*, where the connectivity is shared by all the frames and implicitly captures *inter-frame surface correspondence*. Not only is such a representation more efficient for storage and transmission, because of the shared connectivity, it is also much more convenient for processing, since common procedures, such as texturing, editing or movement analysis can exploit the known surface correspondence. On

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9

the other hand, current state of the art approaches to converting a TVM into a dynamic mesh suffer from many problems and work robustly only under rather constraining conditions.

Working directly with TVMs provides a much greater versatility, however, for many tasks, such as compression, time-consistent editing and others, it is necessary to build an *auxiliary model* that captures the temporal correspondence that is present in the data. It has been observed previously that it is often difficult to establish correspondence of surface elements, since such correspondence loses bijectivity even in the common case of self-contact of objects in the input data. Volume correspondence, on the other hand, is bijective for a wider variety of possible inputs, limited by the requirement of approximately constant overall volume. Therefore, some recent methods have focused on establishing correspondence of volume elements by means of tracking, achieving partial success and applicability in certain scenarios.

One of the drawbacks of the current approaches is that they rely on a frame-by-frame processing procedure, gathering information about the nature of the objects captured in the data in chronological order. This approach prevents information from frames that appear later in the sequence to influence the tracking results (and the induced correspondence information) of preceding frames. This leads to certain artifacts in the tracking results, which in turn hinder application of the tracking in scenarios that are sensitive to tracking errors, such as compression or time-consistent editing.

In particular, when tracking volume elements through time, it is necessary to capture information on which elements are tightly bound together - in previous works, this binding has been termed **affinity**. When processing the sequence forwards in time, the affinity gets constantly updated, as parts of the objects separate or come into contact. The updating may in turn lead to volume elements transitioning between separate components (we refer to such centers as *irregular*), due to temporary self contact, as illustrated in Fig. 1. Also, even when the qualitative change of separating a volume element from its component does not occur, the continuously updated affinity constantly lags behind the actual shape changes, leading to sub-optimal tracking results.

In this paper, we address these issues by proposing an improved tracking procedure, which efficiently eliminates most of the problems encountered previously. Our main contributions are:

- an improved approach to evaluating volume element affinity, which eliminates reconnecting of previously separated components caused by the infinite impulse response (IIR) filter used in the state of the art,
- a robust measure capable of identifying incorrectly tracked volume elements, which allows removing them from the intermediate result,
- a new post-processing phase that optimises tracking criteria globally, taking the whole sequence into account, allowing temporal propagation of information in both directions.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9

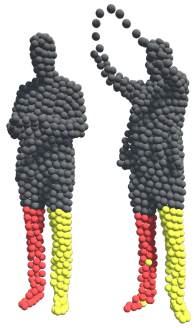


Fig. 1: Example of irregularly tracked centers. Consistent colouring of legs of the subject reveals two yellow coloured centers transitioning to the red coloured part.

The global optimisation step allows for correcting the tracking imperfections caused by the removal of incorrectly tracked elements. We demonstrate the superiority of the proposed tracking strategy both quantitatively and qualitatively.

2 Related work

The most popular way of obtaining a temporal model to represent TVMs is surface tracking. Usually, a certain template surface is sequentially aligned to all the frames using non-rigid registration [15,13]. The simplest methods rely on the template surface given prior and an assumption that it reflects the ground-truth topological information. Such methods usually fail in the presence of frequent erroneous self-contact in the input. This issue might be mitigated to some extent by subdividing the surfaces into patches [5,11], or by identifying frames with significant change in appearance and working with subsequences between such frames [9,6,17,14]. Bojsen-Hansen et al. [2] were able to detect topology changes and adjust the template shape accordingly, however, they incorrectly assume surface correspondences to be bijective outside of the adjusted parts. Budd et al. [4] build a shape similarity tree, which allows alignment of the more similar rather than subsequent frames.

In the presence of self-contacts, the volume correspondences are more likely bijective than the surface correspondences. This has been already utilised by Huang et al. [10,12], who proposed non-rigid registration of centroidal Voronoi tessellations (CVTs). However, their approach does not consider that volume elements move coherently together. Dvořák et al. [7] proposed to track volume elements called *centers*, which are not inherently CVTs, but instead are regularised to achieve coherence of the movement, considering a spatial neighbourhood of each center. This approach was improved [8] by introducing a more appropriate notion of center neighbourhood based on similarity of motion in already tracked frames and motion regularisation, which works better for rigidly moving parts.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9

Since our proposed approach builds on this method, it will be described in more detail in Section 3.

Recently, machine learning models became popular for representing temporal sequences. These are especially successful on sparse data (e.g., single-view RGBD video). Relevant to our work is, for example, OccupancyFlow [16], a learned occupancy function deformed by a neural vector field, as well as the work of Božič et al. [3] who train a neural deformation graph. The main limitation of neural models is, however, that working with them is less intuitive and thus their application is limited for example in time-consistent editing.

The rest of the paper is organised as follows. First, we will describe the key concepts of volume tracking relevant to this paper. In Section 4, we formulate improved affinity weights based on the maximum distances of centers and motion dissimilarities encountered in the already processed frames. These weights are a drop-in replacement for the original IIR filter based affinity. While significantly reducing the tracking error, these weights still do not prevent the occurrence of irregular centers. To this end, in Section 5 we discuss an irregularity measure based on the distance to the trajectory of the nearest center, which allows identifying centers to be removed from the tracking results. In Section 6, we discuss a postprocessing of the centers to attenuate the influence of irregular center removal.

3 As-rigid-as-possible volume tracking

In this section, we briefly review the principles of a state-of-the-art method for volume tracking of TVMs, focusing on parts that are relevant to this paper. For a full description of the method, we refer the reader to the original paper [8]. The input to the method is a sequence of triangle meshes, denoted frames, with no assumption on the coherence of their connectivity. The method finds a fixed set C of N points (denoted centers), each representing a small volume surrounding it, whose positions vary in time. Each center follows a certain trajectory $\mathbf{c}_i = \left[\mathbf{c}_i^{(0)}, \mathbf{c}_i^{(1)}, \ldots, \mathbf{c}_i^{(F-1)}\right] \in \mathbb{R}^{3F}$, where F is the number of frames and $\mathbf{c}_i^{(f)}$ is the position of the i-th center in the f-th frame. The method aims to uniformly distribute the centers inside the enclosed volume of each frame, while ensuring that each center moves coherently with its neighbouring centers.

First, each frame is converted into a dense regular square voxel grid by sampling the indicator function $IF(\mathbf{x})$, which returns 1 in the interior and 0 otherwise. Alternatively, the method can also accept the sequence of voxel grids directly as input, which means that it can be applied to any sequence of shapes for which it is possible to determine the inside/outside information with acceptable amount of certainty (e.g., implicit representations, point clouds, etc.).

Center positions in the first frame are obtained by sampling n random occupied voxels and uniform distribution of centers is achieved by the Lloyd's algorithm: For each center, its Voronoi cell V_i^0 of occupied voxel positions is

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9

iteratively evaluated, such as

$$V_i^{(f)} = \left\{ \mathbf{x} : IF(\mathbf{x}) = 1 \wedge \|\mathbf{x} - \mathbf{c}_i^{(f)}\| \leq \|\mathbf{x} - \mathbf{c}_j^{(f)}\| \right\},$$

for every j, and the center is moved to the centroid $\bar{\mathbf{x}}_i^{(0)}$ of such cell.

For each subsequent frame, the method first obtains an initial distribution of the centers by linear extrapolation from the previous frame. Then, the positions are adjusted in an optimisation process, in which a tracking energy $E = E_s + \beta E_u$ is minimised, where β is a weighting constant ($\beta = 1$ by default). The uniformity energy term E_u enforces uniform distribution of the centers inside the volume. It is formulated as a sum of squared distances between the centers and the centroids of their corresponding Voronoi cells:

$$E_u = \frac{1}{2} \sum_{\mathbf{c}_i \in C} \|\mathbf{c}_i^{(f)} - \bar{\mathbf{x}}_i^{(f)}\|^2.$$

The smoothness of the movement measured in E_s is evaluated as

$$E_s = \frac{1}{2} \sum_{\mathbf{c}_i \in C} \|\mathbf{c}_i^{(f)} - \mathbf{p}_i^{(f)}\|^2,$$

where $\mathbf{p}_i^{(f)}$ is a prediction of the center position obtained using rigid transformation estimated from the movement of neighbouring centers and affinity weights from the previous frame $w^{(f-1)}$:

$$\mathbf{p}_i^{(f)} = \mathcal{A}_{i|w^{(f-1)}}^{(f)}(\mathbf{c}_i^{(f-1)}) = \mathbf{R}_{i|w^{(f-1)}}^{(f)} \mathbf{c}_i^{(f-1)} + \mathbf{t}_{i|w^{(f-1)}}^{(f)}.$$

Considering center positions fixed, the rigid transformation $\mathcal{A}_{i|w}^{(f)} = (\mathbf{R}_{i|w}^{(f)}, \mathbf{t}_{i|w}^{(f)})$ at a frame f given a certain set of weights w can be found minimising

$$(\mathbf{R}_{i|w}^{(f)}, \mathbf{t}_{i|w}^{(f)}) = \underset{\mathbf{R} \in \mathrm{SO}(3), \mathbf{t} \in \mathbb{R}^3}{\arg\min} \sum_{w(i,j) \geq \mu} w(i,j) \left\| \mathbf{c}_j^{(f)} - (\mathbf{R}\mathbf{c}_j^{(f-1)} + \mathbf{t}) \right\|^2,$$

where $\mu = 0.001$ is a threshold parameter to speedup the computation process by considering only relevant weights. Such transformation can be found in closed form using singular-value decomposition [18].

To optimise the energy E, the method interleaves between calculating the predictions $\bar{\mathbf{x}}_i^{(f)}$ and $\mathbf{p}_i^{(f)}$ with fixed positions of centers and then updating the positions with fixed predictions:

$$\mathbf{c}_i^{(f)} = \frac{\mathbf{p}_i^{(f)} + \beta \bar{\mathbf{x}}_i^{(f)}}{1 + \beta}.$$

The optimisation process is terminated when the change in $\mathbf{c}_i^{(f)}$ is sufficiently small or a fixed number of iterations has been reached.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9

Once the final positions in the current frame are obtained, the affinity weights are updated, so that they reflect the observed changes in the relations between centers in the currently processed frame. The method considers center relations based on spatial proximity in a single frame

$$a_p^{(f)}(i,j) = \exp(-\sigma_p \cdot \left\| \mathbf{c}_i^{(f)} - \mathbf{c}_j^{(f)} \right\|^2),$$

where σ_p is a parameter controlling the width of the Gaussian function.

The method also attempts to separate topologically distant parts by combining the spatial proximity with motion dissimilarity that is measured as:

$$a_m^{(f)}(i,j) = \exp(-\sigma_m \cdot d_i^{(f)}(\mathcal{A}_{i|w^{(f-1)}}^{(f)}, \mathcal{A}_{j|w^{(f-1)}}^{(f)})^2),$$

where σ_m is another Gaussian width parameter and $d_i^{(f)}(\mathcal{A}, \mathcal{B})$ is the distance between two rigid transformations that is evaluated by measuring distances of points around $\mathbf{c}_i^{(f)}$ (voxel positions in $V_i^{(f)}$ in our case) transformed by both $\mathcal{A}$ and $\mathcal{B}$

$$d_i^{(f)}(\mathcal{A}, \mathcal{B}) = \frac{1}{|V_i^{(f)}|} \sum_{\mathbf{v}_k \in V_i^{(f)}} \left\| \mathcal{A}(\mathbf{v}_k) - \mathcal{B}(\mathbf{v}_k) \right\|.$$

Instead of setting the Gaussian width parameters σ directly, they are calculated from parameters ρ with clearer geometric meaning: $\sigma = -\ln(0.5)/\rho^2$, which determine at which distance the Gaussian function drops to 0.5.

To propagate the already observed information throughout the sequence, an IIR filter with falloff parameter α is applied on motion dissimilarity:

$$a_{\mathrm{IIR}}^{(f)}(i,j) = \alpha a_m^{(f)}(i,j) + (1-\alpha) a_{\mathrm{IIR}}^{(f-1)}(i,j).$$

Finally, the spatial proximity $a_p^{(f)}(i,j)$ is combined with the IIR filtered motion dissimilarity $a_{\mathrm{IIR}}^{(f)}(i,j)$ to form the weights $w^{(f)}(i,j)$ that will be used to optimise the positions in the next frame:

$$w^{(f)}(i,j) = a_p^{(f)}(i,j) \cdot a_{\mathrm{IIR}}^{(f)}(i,j). \tag{1}$$

With this knowledge, we can proceed with discussing the contributions of this paper.

4 Maximum distance based affinity

Similarly to the original weight formulation in Eq. 1, the new weight is also computed as a product of spatial proximity and motion dissimilarity:

$$\tilde{w}^{(f)}(i,j) = \tilde{a}_p^{(f)}(i,j) \cdot \tilde{a}_m^{(f)}(i,j).$$

The difference is how the center proximity $\tilde{a}_p^{(f)}(i,j)$ and the motion dissimilarity $\tilde{a}_p^{(f)}(i,j)$ are formulated.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9

The previous method measured the spatial center proximity using the Euclidean distance of centers in a single frame, ignoring the information from previous frames. The main limitation of such approach is the fact that as two topologically distant or separated parts come to near proximity, the affinity between their centers increases. Assuming the tracked sequence represents piecewise rigid objects, it could be more appropriate to use geodesic distance inside the volume, which, ideally, should be roughly constant throughout the sequence. However, due to self contact present in real-world data, the topological information in a given frame might be incorrect, resulting in introduction of erroneous decrease in such a measure. Additionally, geodesic distance is computationally expensive to evaluate at the frequency required by the tracking pipeline. We instead propose to approximate this quantity by the largest Euclidean distance encountered in all frames up to and including the current frame:

$$\tilde{a}_p^{(f)}(i,j) = \exp\left(-\sigma_p \cdot \max_{0\le k\le f} \left\|\mathbf{c}_i^{(f)} - \mathbf{c}_j^{(f)}\right\|^2\right).$$

When examining relative positions of a certain pair of centers in time, we observe that the Euclidean distance between them fluctuates, but it is never larger than their geodesic distance. Note that our goal is not to evaluate this quantity precisely, but to correctly differentiate between the true connected neighbours of a center and the topologically distant centers in near proximity (see Figure 2).

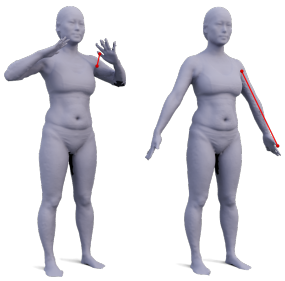


Fig. 2: Spatial proximity in a single frame might not reflect the underlying topology of the represented object. Left: Two topologically distant points in near proximity. Right: Examining a different frame reveals that they should not be considered as neighbouring/affine.

An analogous observations can be made about the similarity of the movement. If a pair of centers moved significantly differently in the past, then they cannot both belong to the same rigid part, even when the movement has been almost identical in several previous frames. Instead of an IIR filter, we thus propose to also use the maximum dissimilarity over all already processed frames:

$$\tilde{a}_m^{(f)}(i,j) = \begin{cases} 1, & f = 0 \\ \exp\left(-\sigma_m \cdot \max_{1\le k\le f} d_i^{(f)}(\mathcal{A}^{(f)}_{i|\tilde{w}^{(f-1)}}, \mathcal{A}^{(f)}_{j|\tilde{w}^{(f-1)}})^2\right), & \text{otherwise} \end{cases}.$$

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9

In the first frame, we have no information about the movement, therefore we assume there is no difference and instead solely rely on $\tilde{a}_p^{(0)}(i,j)$ when computing the affinity weights.

Setting Gaussian widths σ_m and σ_p (resp. ρ_m and ρ_p) to obtain satisfactory results is a task specific to the scale of the data and complexity of the motion. In our experiments, we have obtained best results with $\rho_p = \rho_m = 0.125$ for human performance capture. For synthetic datasets, where the bounding box was significantly larger and the motion was mainly rigid, we have determined that best results were obtained with $\rho_p = 0.2$ and $\rho_m = 0.05$.

The previous IIR-filter-based affinity depends on a falloff parameter α, which controls how the affinity reacts to occurring changes. Setting α too low results in slow reactions. On the other hand, too high α means that the affinity forgets faster the separation that occurred in the past. Our new formulation of affinity reacts more dynamically to changes and also reflects every observed separation.

5 Irregular center detection

Once the frame-by-frame tracking is finished (regardless of the affinity weights used), we can analyse the achieved results and detect the *irregular centers*. To this end, we evaluate an irregularity measure $I_i = \min_j \|\mathbf{c}_i - \mathbf{c}_j\|_2^2$, where $\mathbf{c}_i$ is the center trajectory and $\|\cdot\|_2^2$ is the squared Euclidean norm. If a center is correctly tracked, there should exist another center with a similar trajectory in the near proximity. Since an irregular center changes suddenly its relative position to its neighbouring centers, even the distance to the closest center to its trajectory is expected to be higher than for the correctly tracked centers (see Figure 3).

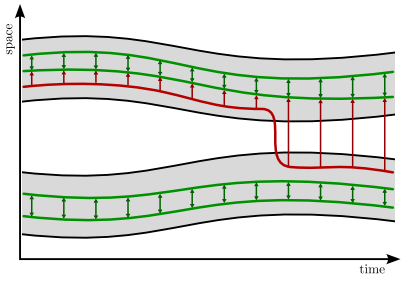


Fig. 3: Irregular center detection using distance to closest trajectory. Arrows indicate distances that contributed to the computation. Red trajectory has a much higher I_i and is correctly detected as irregular.

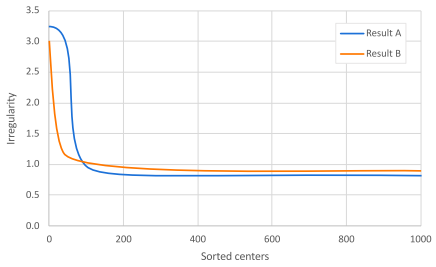


Fig. 4: Example comparison of two irregularity curves. The Result B outperforms the Result A as its curve drops faster to satisfying values of I_i.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9

The value of I_i must be considered in the context of the values of all centers, as it depends on various factors, e.g., center count, scale of the data and the dynamics of the movement. We can also use this measure to quantify the success of tracking in terms of the presence of irregular centers, by sorting all the values in descending order and plotting them as a curve. By comparing the curves resulting from different tracking methods, we can determine which results are less influenced by the presence of irregular centers (see Figure 4), as long as the results were tracked in the same input sequence and the center count is similar (although not necessarily equal). When attempting to improve the tracking results, one of our goals is to narrow or eliminate the part of the curve with I_i significantly higher than the correctly tracked centers, while not significantly increasing the irregularity of such centers.

6 Global optimisation

Simply removing a certain number of centers with the highest I_i actually does not lead to an improvement in terms of flattening the irregularity curve. The uniformity of the centers distribution is violated, since removed centers leave an uncovered volume. Re-running the frame-by-frame tracking with the irregular centers removed might still not prevent new irregular centers from appearing, even when the final affinity weights obtained in the initial tracking are utilised. Instead, we propose to follow the irregular center removal with adjustment of the remaining tracked trajectories of centers in a global optimisation process.

6.1 Global tracking energy

The objectives of the global optimisation are identical to the frame-by-frame tracking. We optimise a global energy $\hat{E}$ consisting of uniformity and motion smoothness energy terms $\hat{E} = \hat{E}_s + \hat{\beta}\hat{E}_u$.

The uniformity term is the same as in the frame-by-frame tracking, except for that it is evaluated for all the frames in the sequence at once:

$$\hat{E}_u = \frac{1}{2} \sum_{\mathbf{c}_i \in C} \sum_{f=0}^{F-1} \|\mathbf{c}_i^{(f)} - \bar{\mathbf{x}}_i^{(f)}\|^2.$$

If we consider the centroids $\bar{\mathbf{x}}_i^{(f)}$ fixed, we can approximate the gradient by these partial derivatives:

$$\frac{\partial \hat{E}_u}{\partial \mathbf{c}_i^{(f)}} \approx \mathbf{c}_i^{(f)} - \bar{\mathbf{x}}_i^{(f)}.$$

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9

The global smoothness energy is evaluated as

$$\hat{E}_s = \frac{1}{2}\left(\sum_{\mathbf{c}_i \in C}\sum_{f=1}^{F-1}\left\|\mathbf{c}_i^{(f)} - \mathbf{p}_i^{(f)}\right\|^2 + \sum_{\mathbf{c}_i \in C}\sum_{f=0}^{F-2}\left\|\mathbf{c}_i^{(f)} - \mathbf{q}_i^{(f)}\right\|^2\right),$$
$$\mathbf{p}_i^{(f)} = \mathcal{A}_{i|\omega}^{(f)}\left(\mathbf{c}_i^{(f-1)}\right),$$
$$\mathbf{q}_i^{(f)} = \mathcal{A}_{i|\omega}^{(f+1)^{-1}}\left(\mathbf{c}_i^{(f+1)}\right),$$

where $\mathbf{p}_i^{(f)}$ and $\mathbf{q}_i^{(f)}$ are forward and backward rigid motion predictions of the center position at frame f, using rigid transformations estimated given overall movement-based affinity weights ω (see Eq. 2). Considering such predictions fixed, the partial derivatives, which form the approximated gradient, are as follows:

$$\frac{\partial \hat{E}_s}{\partial \mathbf{c}_i^{(f)}} \approx \begin{cases} \mathbf{c}_i^{(f)} - \mathbf{q}_i^{(f)}, & f = 0 \\ \mathbf{c}_i^{(f)} - \mathbf{p}_i^{(f)} - \mathbf{q}_i^{(f)}, & 1 \leq f \leq F-2 \\ \mathbf{c}_i^{(f)} - \mathbf{p}_i^{(f)}, & f = F-1 \end{cases}$$

6.2 Optimisation strategy

The optimisation process is iterative, working with a set of trajectories C, whose initial values are given by the original tracking results with irregular centers removed. In each iteration, we evaluate the energy $\hat{E}(C)$ and the approximated gradient $\nabla\hat{E}$, and construct a candidate set of trajectories $\bar{C}$, where the center positions are calculated as

$$\bar{\mathbf{c}}_i^{(f)} = \mathbf{c}_i^{(f)} - \lambda\left(\frac{\partial \hat{E}_s}{\partial \mathbf{c}_i^{(f)}} + \hat{\beta}\frac{\partial \hat{E}_u}{\partial \mathbf{c}_i^{(f)}}\right).$$

First, lambda is set to $\lambda = 0.1$ and then it is iteratively scaled by $\frac{1}{2}$ until $\hat{E}(\bar{C})$ is smaller than $\hat{E}(C)$, or a specified number of attempts has been reached. If an improvement in terms of energy is achieved, we set $C = \bar{C}$ and continue to the next iteration. Otherwise, the process is terminated and C is the resulting set of trajectories. The optimisation process can also be terminated after a specified number of iterations (20 in our experiments).

Such an optimisation strategy does not necessarily converge to a global optimum. If an irregular center was left in the initial set C, the local steps in the gradient direction will not straighten its trajectory in order to eliminate the transition between disconnected components. The locality of the changes is, however, also an advantage, since the local trajectory adjustments ensure that the objectives are met, while not introducing any large sudden changes, and therefore no new irregular centers can appear.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9

6.3 Global movement-based affinity

The affinity utilised in the global optimisation process directly considers only the dissimilarity of motion:

$$\omega(i,j) = \exp\left(-\sigma_{\text{gm}} \cdot \max_{0 \le f < F} d_i^{(f)} \left(\mathcal{A}^{(f)}_{i|\omega_{\max}}, \mathcal{A}^{(f)}_{j|\omega_{\max}}\right)^2\right), \tag{2}$$

where σ_{gm} is a parameter controlling the tolerance of the affinity to dissimilar motions. The overall spatial proximity $\omega_{\max}(i,j)$ of centers is also considered, but only for estimating the transformations $\mathcal{A}^{(f)}_{i|\omega_{\max}}$:

$$\omega_{\max}(i,j) = \exp\left(-\sigma_{\text{gp}} \cdot \max_{0 \le f < F} \left\|\mathbf{c}_i^{(f)} - \mathbf{c}_j^{(f)}\right\|^2\right).$$

The motivation to mainly rely on the motion information instead of using both motion and proximity combined is the following: If a center belongs to a large rigidly moving part of the object, we want it to be influenced by the whole part rather than only a certain small neighbourhood around it. Having all the positions in each frame at hand, we can confidently rely solely on this information without worrying about the rigid part suddenly splitting in a later frame.

7 Experimental results

7.1 Influence of the proposed affinity

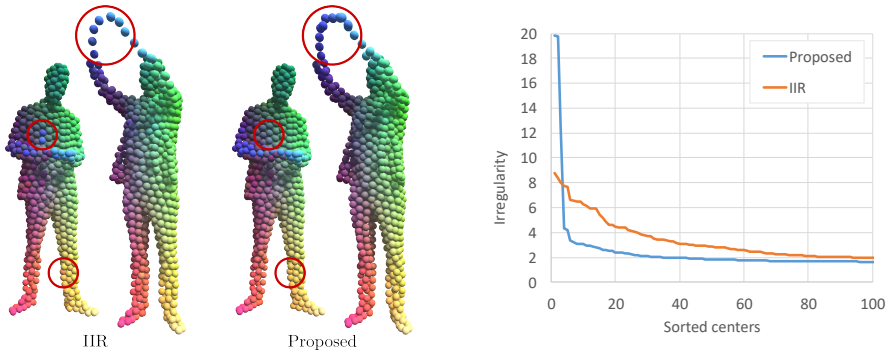
To evaluate how the previous forward tracking pipeline benefits from the proposed maximum distance based affinity, we have compared the new tracking results with those reported previously [8] using default configurations for both methods. The comparison included all the previously studied datasets (including selected sequences from D-FAUST dataset [1]) except for *pentagonal_prism* and *collision* datasets, which we believe were already tracked correctly with the previous method. The tracking quality was evaluated using the metrics PCAC, which measures the complexity of the tracked trajectories (lower is assumed better, although lowering below a certain threshold given by the true complexity of the movement is not desirable) and DFU, which measures relative standard deviation of Voronoi cell sizes (a lower value indicates a more uniform covering). For details on the measures, see the original paper [8]. The results are shown in Table 1.

Incorporating the newly proposed affinity results in a considerable improvement over the original affinity on all the sequences. Visually, the results contain less irregular centers, which can be seen when assigning each center a consistent color by interpreting the first three PCA coefficients of its trajectory as RGB values, and the centers achieve better coverage over problematic parts (see Figure 5a). This is also reflected by the irregularity curves (see Figure 5b).

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9

Table 1: Comparison of forward-tracked results using proposed and original (IIR) affinity. Highlighted are the best results for a given dataset.

	F	Proposed		IIR	
		PCAC	DFU	PCAC	DFU
gears	60	**1.785**	**0.070**	1.816	0.071
casual_man	545	**10.587**	**0.127**	12.289	0.206
samba	175	**5.060**	**0.215**	5.547	0.262
DF_50020_knees	515	**5.117**	**0.111**	6.764	0.179
DF_50009_chicken_wings	212	**2.855**	**0.116**	3.583	0.155
DF_50004_jumping_jacks	360	**6.040**	**0.130**	7.826	0.175



(a) Visual comparison. (b) Largest 100 values of I_i.

Fig. 5: Tracking results for *casual_man* dataset.

7.2 Irregular center removal

In this experiment, we have studied the effects of various strategies for removing 20 irregular centers from tracking results obtained through forward tracking with our proposed affinity on the *casual_man* dataset from Section 7.1. The strategies differed in the number of global optimisations performed n_{go} and in number of removed centers each optimisation n_{rem}. Parameters of the global optimisation were as follows: $\rho_{\text{gp}} = 0.125, \rho_{\text{gm}} = 0.03$ and $\hat{\beta} = 0.5$. Note that these values were selected empirically and slightly different values yield similar results. Table 2 shows the measured PCAC and DFU values and the irregularity curves are shown in Fig. 6. For comparison, we also include results for center removal without global optimisation.

With growing n_{go}, the improvement process achieves better coverage of volume, which is reflected in the DFU measure. However, we can also see a negative trend in terms of irregularity. Best values of PCAC were achieved with $n_{\text{go}} = 5$. This is also reflected in a visual inspection of results (see Fig. 7). Fig. 7a shows centers that were detected as irregular. It can be seen that the detected centers indeed travel across different body parts. The increased irregularity with growing

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9

Table 2: Comparison of PCAC and DFU measures for various irregular center removal strategies on *casual_man* dataset forward tracked using our proposed affinity.

n_{rem}	20	20	4	1
n_{go}	—	1	5	20
PCAC	10.154	8.760	**8.578**	8.869
DFU	0.160	0.149	0.144	**0.134**

n_{go} is reflected by certain number of centers oscillating to cover a larger volume, which is unfortunately visible only when the tracked centers are animated.

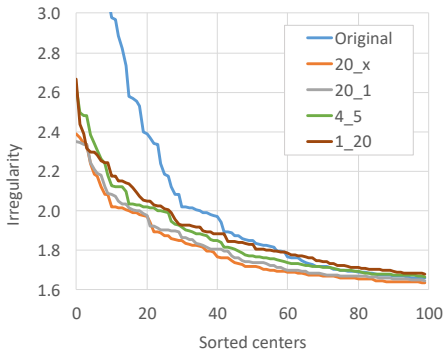


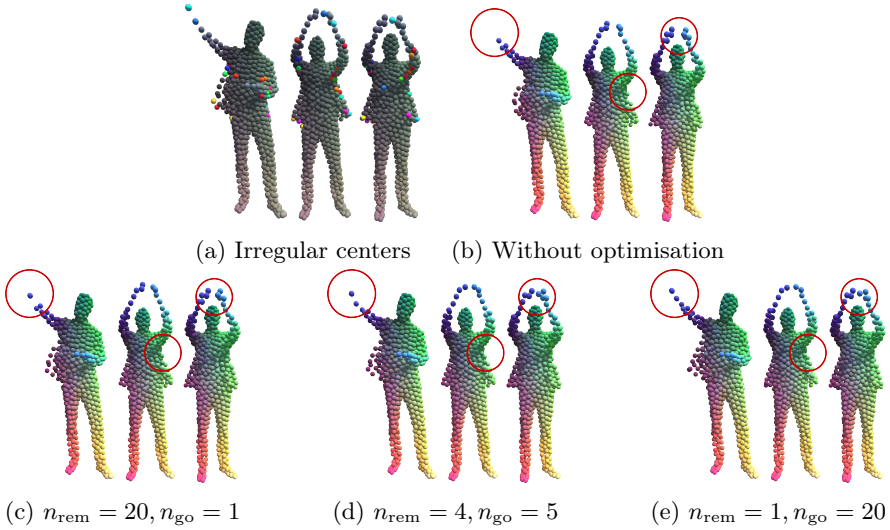
Fig. 6: Comparison of first 100 I_i after center removal.

8 Conclusions

In this paper, we have shown that the forward volume tracking results can be further considerably improved by incorporating volume element filtering followed by a global optimisation, which assures volume coverage without introducing new irregularly tracked centers. The experiments show that it is beneficial to remove centers in small batches, rather than all at once or each center separately. The novel affinity for the forward tracking method, which we also proposed, considerably reduces tracking imperfections, which is reflected in all considered tracking quality metrics.

Our approach shares the limitations of the previous volume tracking methods, as it also cannot handle sequences without sufficient notion of inside/outside information. Additionally, it might be more sensitive to noise, since any error introduced in computing the affinity might result in incorrect split of two affine

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9



(a) Irregular centers (b) Without optimisation

(c) $n_{\text{rem}} = 20, n_{\text{go}} = 1$ (d) $n_{\text{rem}} = 4, n_{\text{go}} = 5$ (e) $n_{\text{rem}} = 1, n_{\text{go}} = 20$

Fig. 7: Results for various irregular center removal strategies for *casual_man* dataset with target of 20 centers to be removed. Highlighted are the areas with the most notable differences.

centers. In forward tracking, this affects only the frames after the occurrence of the error. However, in global optimisation, the whole sequence is influenced.

In the future, we would like to study the means of inserting centers into the intermediate tracking results as a complementary process to filtering and global optimisation, in order to further improve the coverage of the tracked volume. We also believe that the volume tracking would benefit from incorporating a notion of a particular shape associated with each center, instead of representing it as a discrete point. A reference implementation of the algorithm is available at https://gitlab.kiv.zcu.cz/jdvorak/arap-volume-tracking.

Acknowledgement

This work was supported by the project 20-02154S of the Czech Science Foundation. Jan Dvořák and Filip Hácha were partially supported by the University specific research project SGS-2022-015, New Methods for Medical, Spatial and Communication Data. The authors thank Diego Gadler from AXYZ Design, S.R.L. for providing some of the test data.

References

1. Bogo, F., Romero, J., Pons-Moll, G., Black, M.J.: Dynamic FAUST: Registering human bodies in motion. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Jul 2017)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9

2. Bojsen-Hansen, M., Li, H., Wojtan, C.: Tracking surfaces with evolving topology. ACM Trans. Graph. **31**(4) (Jul 2012)
3. Božič, A., Palafox, P., Zollhöfer, M., Dai, A., Thies, J., Nießner, M.: Neural deformation graphs for globally-consistent non-rigid reconstruction. arXiv preprint arXiv:2012.01451 (2020)
4. Budd, C., Huang, P., Klaudiny, M., Hilton, A.: Global non-rigid alignment of surface sequences. International Journal of Computer Vision **102**(1-3), 256–270 (2013)
5. Cagniart, C., Boyer, E., Ilic, S.: Free-form mesh tracking: A patch-based approach. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1339–1346 (2010). https://doi.org/10.1109/CVPR.2010.5539814
6. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Trans. Graph. **34**(4) (Jul 2015). https://doi.org/10.1145/2766945
7. Dvořák, J., Vaněček, P., Váša, L.: Towards understanding time varying triangle meshes. In: Paszynski, M., Kranzlmüller, D., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M. (eds.) Computational Science – ICCS 2021. pp. 45–58. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-77977-1_4
8. Dvořák, J., Káčereková, Z., Vaněček, P., Hruda, L., Váša, L.: As-rigid-as-possible volume tracking for time-varying surfaces. Computers & Graphics **102**, 329–338 (2022). https://doi.org/10.1016/j.cag.2021.10.015
9. Guo, K., Xu, F., Wang, Y., Liu, Y., Dai, Q.: Robust non-rigid motion tracking and surface reconstruction using l0 regularization. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 3083–3091 (2015). https://doi.org/10.1109/ICCV.2015.353
10. Huang, C.H., Allain, B., Franco, J.S., Navab, N., Ilic, S., Boyer, E.: Volumetric 3d tracking by detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
11. Huang, C.H., Boyer, E., Ilic, S.: Robust human body shape and pose tracking. In: 2013 International Conference on 3D Vision - 3DV 2013. pp. 287–294 (2013). https://doi.org/10.1109/3DV.2013.45
12. Huang, C.H.P., Allain, B., Boyer, E., Franco, J.S., Tombari, F., Navab, N., Ilic, S.: Tracking-by-detection of 3d human shapes: From surfaces to volumes. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(8), 1994–2008 (2018). https://doi.org/10.1109/TPAMI.2017.2740308
13. Li, H., Adams, B., Guibas, L.J., Pauly, M.: Robust single-view geometry and motion reconstruction. ACM Trans. Graph. **28**(5), 1–10 (Dec 2009). https://doi.org/10.1145/1618452.1618521
14. Moynihan, M., Ruano, S., Pages, R., Smolic, A.: Autonomous tracking for volumetric video sequences. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1660–1669 (January 2021)
15. Myronenko, A., Song, X.: Point set registration: Coherent point drift. IEEE transactions on pattern analysis and machine intelligence **32**(12), 2262–2275 (2010)
16. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Occupancy flow: 4d reconstruction by learning particle dynamics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
17. Prada, F., Kazhdan, M., Chuang, M., Collet, A., Hoppe, H.: Spatiotemporal atlas parameterization for evolving meshes. ACM Trans. Graph. **36**(4) (Jul 2017). https://doi.org/10.1145/3072959.3073679
18. Sorkine-Hornung, O., Rabinovich, M.: Least-squares rigid motion using svd (2016), technical note

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_9